Laajojen historiallisten aineistojen louhinnan mahdollisuudet ja pullonkaulat

FinELibin aineistopäivä 16.4.2015
Tämä esitys: http://j.mp/FinELib

Mikko Tolonen ja Eetu Mäkelä

Nineteenth Century
Collections Online

Searching 2 of 2 Archives

Search

Advanced Search

Archives available in your library

Religion, Society, Spirituality, and Reform

Science, Technology, and Medicine: 1780-1925

Digital humanities

- DH1.0 ja “uusi aalto”
- Laskennallisten ja tilastollisten menetelmien innovatiivinen käyttö
- Uusia työkaluja “ikuisten” kysymysten tutkimiseen.
- Tutkimuskysymykset edellä!
- Uuden digitointi hyvästä, mutta meillä on jo myös mielenkiintoista aineistoa tutkittavaksi, kunhan sitä edes yritettäisiin käyttää.

Avoimen kulttuuriaineiston vyöry

- Suuret integraattorit suosivat CC0-lisenssiä & julkaisevat kokoamansa datan: Europeana, Digital Public Library of America, The European Library, ~~Finna~~
- Maailman kansalliskirjastot siirtymässä avoimen linkitetyn tiedon yhteistyöhön: Library of Congress, Deutsche Nationalbibliothek, British Library, Bibliothèque nationale de France, ~~Kansalliskirjasto~~
- Museot, Galleriat ja Arkistot seuraavat perässä: British Museum, Kansallisgalleria, …
- Aineiston yhdistämiseen tarvittavia jaettuja auktoriteettikantojakin löytyy: KOKO, VIAF, CIDOC-CRM, Getty AAT, TGN, ULAN, CONA, Iconclass, Wikidata, DBPedia, GeoNames, Pleiades, ...

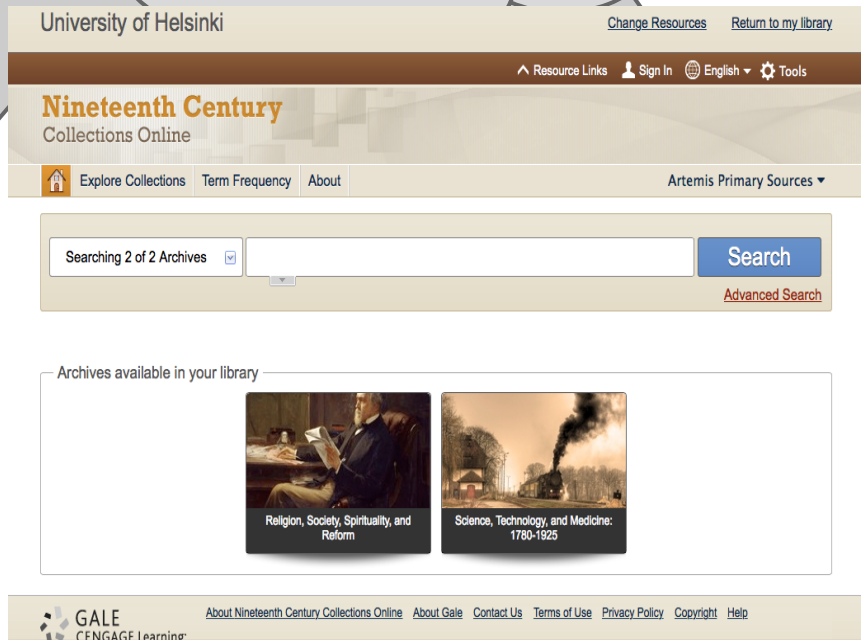
Mikä ei ole avointa?

- Tieteelliset kirjastot tottuneet tuottamaan palveluja yhteistyössä kirjastopalveluyritysten kanssa (EBSCO Information Services, ProQuest LLC, Gale Cengage Learning)
- Usein sisältöpalvelutkin tuotetaan yhdessä, jolloin kirjastopalveluyritys pistää ne lukkojensa taakse:
  - Early English Books Online (ProQuest)
  - Eighteenth Century Collections Online (Gale)
  - Nineteenth Century Collections Online (Gale)
  - State Papers Online (Gale)


painettu ja sähköinen

painettu ja sähköinen

=

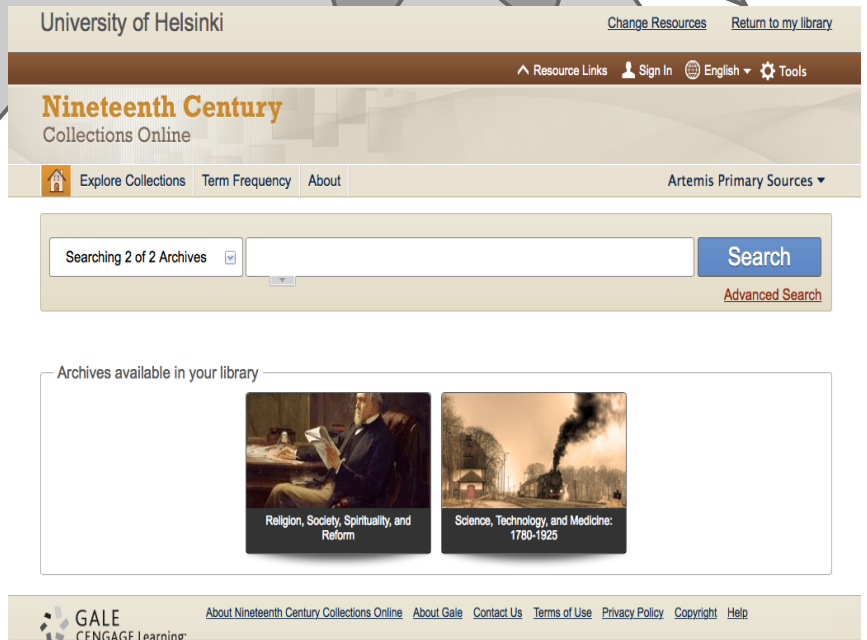


```
<lg type="sestet" rhyme="abababab">
    <l>This was the measure of my soul's <rhyme label="a" corresp="#F">delight</rhyme>;</l>
    <l>It had no power of joy to fly by <rhyme label="b" corresp="#A">day</rhyme>,</l>
    <l>Nor part in the large lordship of the <rhyme label="a" corresp="#D">light</rhyme>;</l>
    <l>But in a secret moon-beholden <rhyme label="b" corresp="#C">way</rhyme></l>
    <l>Had all its will of dreams and pleasant <rhyme label="a" corresp="#B">night</rhyme>,</l>
    <l>And all the love and life that sleepers <rhyme label="b" corresp="#E">may</rhyme>.</l>
</lg>
<lg type="sestet" rhyme="abababab">
    <l>But such life's triumph as men waking <rhyme label="a" corresp="#E">may</rhyme></l>
    <l>It might not have to feed its faint <rhyme label="b" corresp="#F">delight</rhyme></l>
    <l>Between the stars by night and sun by <rhyme label="a" corresp="#A">day,</rhyme></l>
    <l>Shut up with green leaves and a little <rhyme label="b" corresp="#D">light;</rhyme></l>
    <l>Because its way was as a lost star's <rhyme label="a" corresp="#C">way,</rhyme></l>
    <l>A world's not wholly known of day or <rhyme label="b" corresp="#B">night.</rhyme></l>
</lg>
```

=

painettu ja sähköinen

=



Avoimen datan periaatteet

- Ei ainoastaan avoin julkaisu – koko tutkimusprosessi on keskeinen ja sen tulisi olla mahdollisimman avoin.
- Metodit, tutkimus, data ja tulokset kaikki avoimiksi.
- Läpinäkyvyys, toistettavuus, informaali yhteistyö, uudet aloitteet
- Pääsy “raakadataan” on institutionaalinen kysymys (ECCO ja EEBO hidasteina)

Extract & Transform

Data is the new oil?

Data is the new bacon?

Data is the new Bacon.

Tiedonlouhinnan vaihtoehtoja aatehistorian näkökulmasta

Laajojen aineistojen tekstilouhinta

- **Tavoite**: ymmärtää käsitteellistä muutosta ja kielen käyttöä
- **Aineistot**: tekstitietokannat (ECCO, EEBO, digitoidut sanomalehdet jne.)
- **Potentiaali**: kvantitatiivinen evidenssi laadullisen tutkimuksen tukena ja uudet havainnot; etäluenta
- **Käytäntö**: alkuperäisaineistot harvoin avoimesti saatavilla tai sidottu rajoitettuihin käyttöliittymiin
- **Skaalautuvuus**: rajattu; data on kooltaan massiivista ja voi vaatia huomattavaa jalostamista; aineistosta ja menetelmistä riippuvainen;
- **Metodologinen näkökulma**: aatehistorian sovelluksia toistaiseksi hyvin vähän

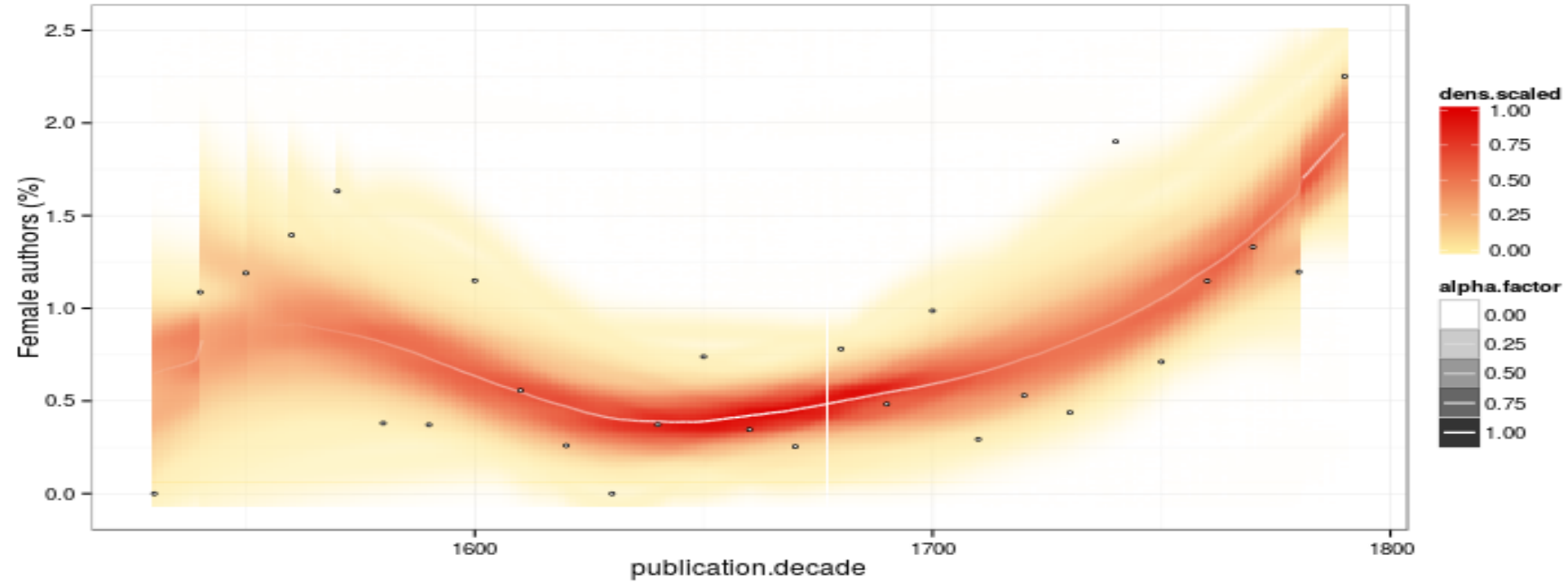
Kuvailutiedot tilastollisena työkaluna

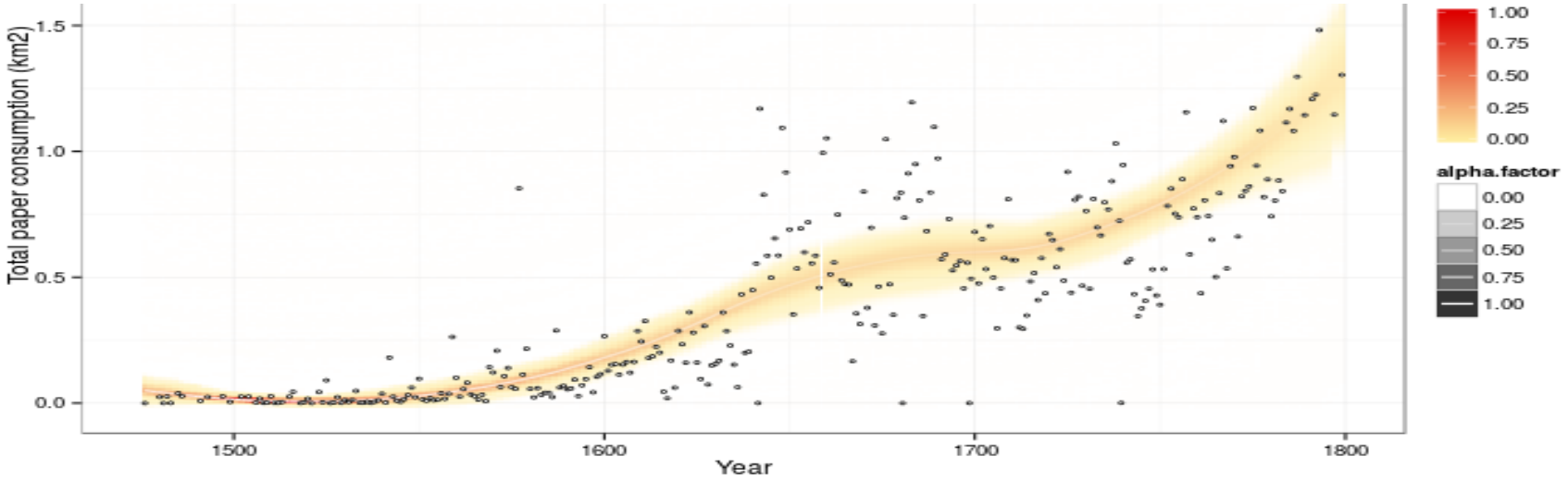
- **Tavoite**: ymmärtää yleisiä trendejä historiallisten ilmiöiden dokumentoinnissa
- **Aineistot**: kirjastot ja arkistot tulvivat erilaisia kuvailutietokantoja (metadata)
- **Potentiaali**: tutkimuskäyttö aliarvioitu
- **Käytäntö**: saatavuudessa ongelmia, mutta vähemmän kuin tekstiaineistojen kohdalla
- **Skaalautuvuus**: erinomainen; data on rakenteista ja sen koko on rajattu
- **Metodologinen näkökulma**: tarjoaa selkeän lähtökohdan uusien käytäntöjen soveltamisessa humanistiseen tutkimukseen.

Metadatan monikäyttöisyys

- Kvantitatiivinen kehys laadulliselle tutkimukselle
- Tiedontuottamisen ymmärtäminen
- Julkaisijat ja niiden verkostot (visualisaatiot)
- Julkaisupaikat, ”cultural transfer”
- ”Historian”, ”filosofian”, ”uskonnon” analysointi (ei genreinä per se), mutta alakategorioina
- Yksittäiset kirjoittajat ja näiden vertailu

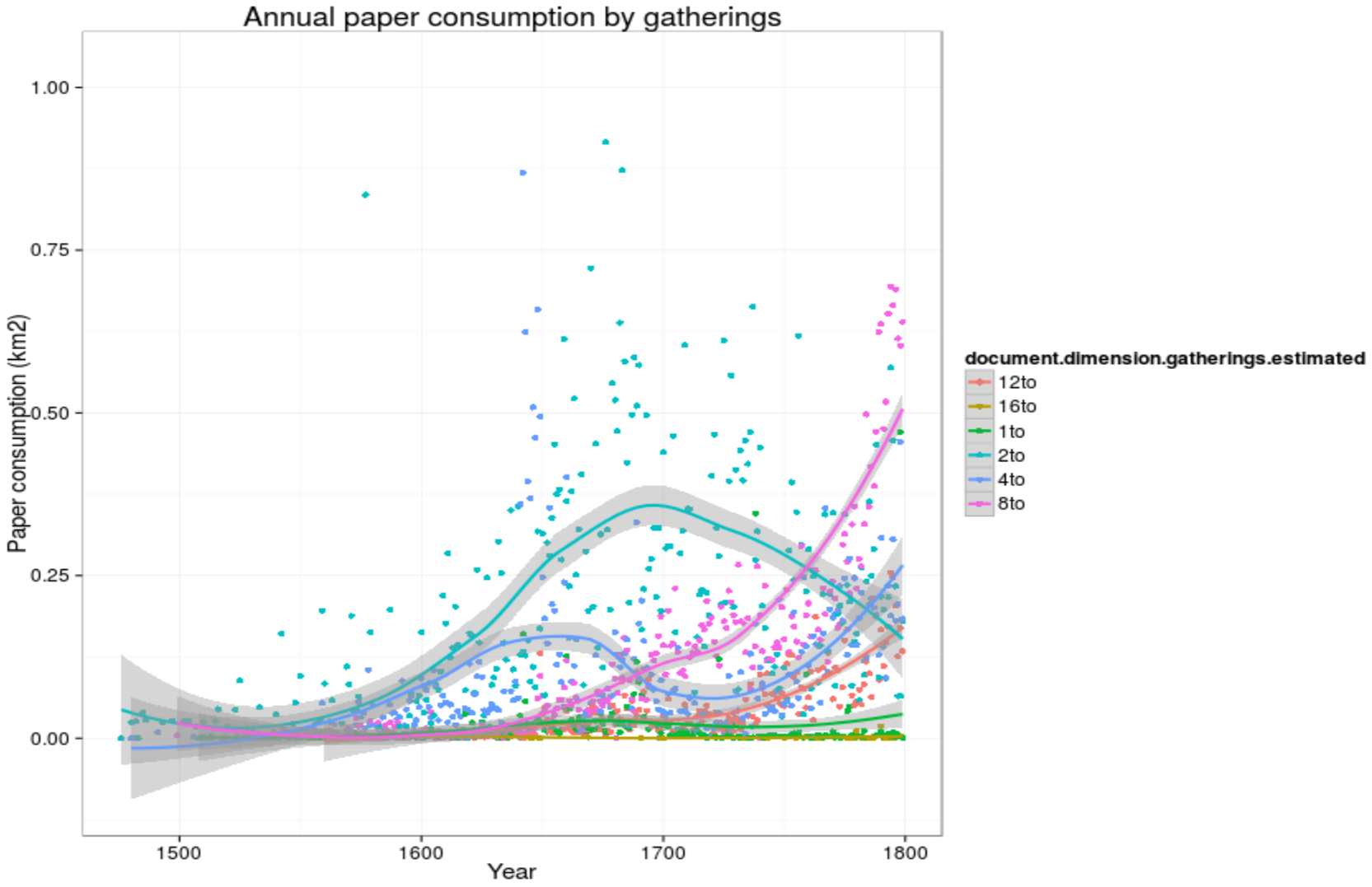
Sukupuolijakaumaa ja paperinkulutusta



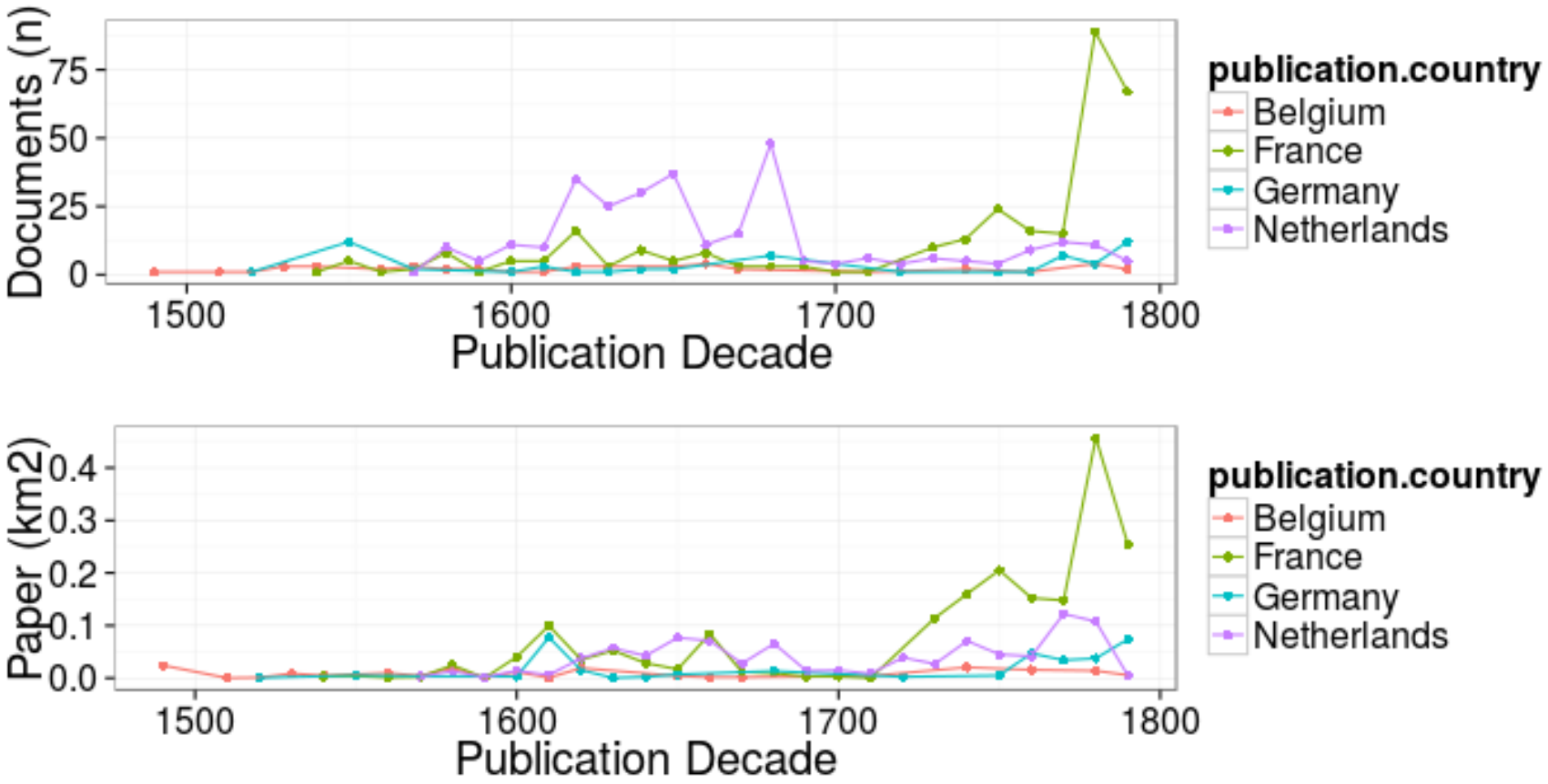


PAGECOUNT-LINKKI

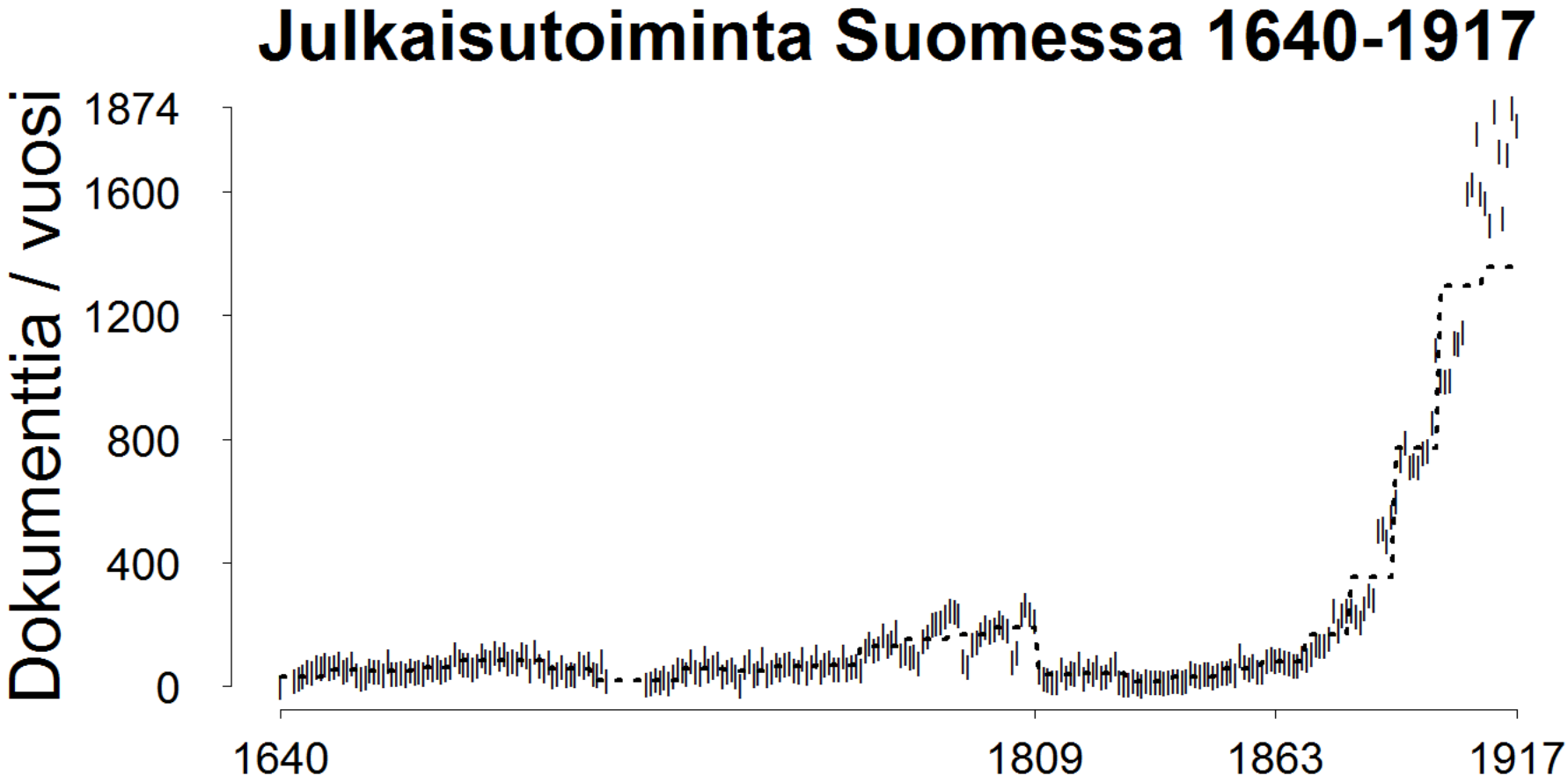
Historia-aineiston kehitystä estc:ssä



Tietoa voi aina syventää



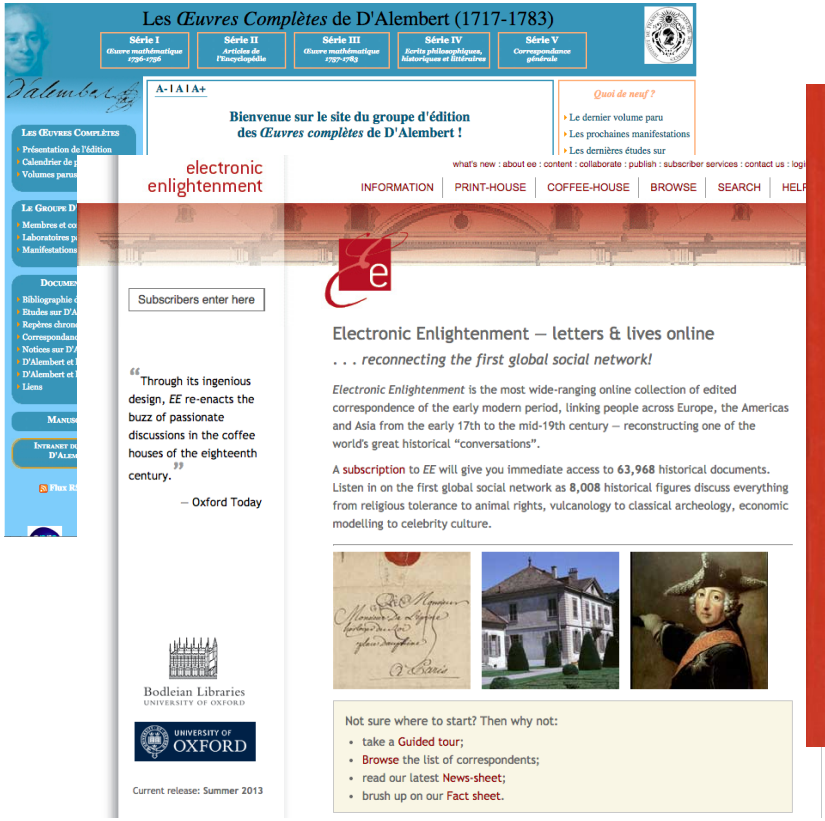
Luetteloinnilla on merkitystä!

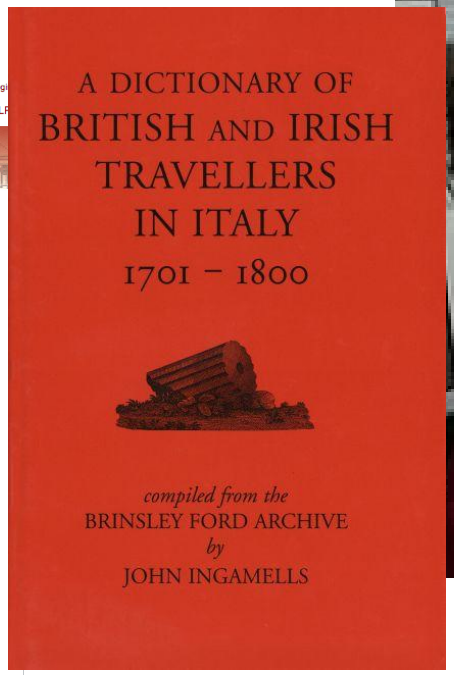


Mitkä asiat yhdistävät Lontoon ja Milanon 1700-luvulla?

Mitkä asiat yhdistävät Lontoon ja Milanon 1700-luvulla?

- Tiedot yhdistetty neljästä hyvin erilaisesta lähteestä:





Mitä voi tehdä metadatalla

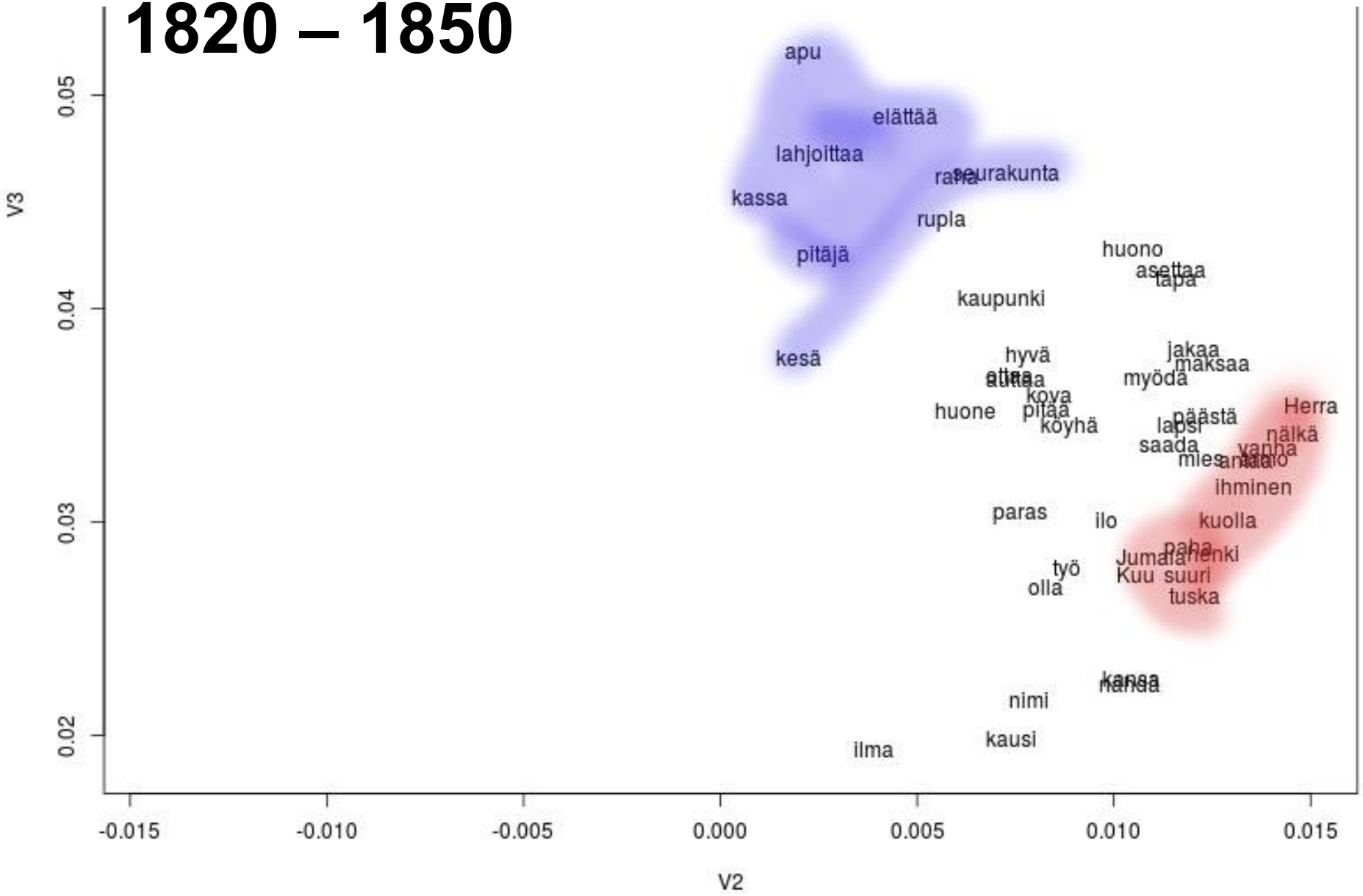
- Electronic Enlightenment + Grand Tour + Procope:
  - Voltairen kirjeenvaihdon maantieteellinen jalanjälki Europeana 4D-työkalussa
  - Voltairen vaikutuspiiri Palladio-työkalussa
- Bibliothèque nationale de France:
  - Missä julkaistaan suhteettoman paljon ranskankielistä filosofiaa 1700-luvulla?
- The Schoenberg Database of Manuscripts:
  - Mistä ovat lähtöisin Sir Thomas Philippsin kokoelman käsikirjoitukset?
  - Kuinka monen käden kautta ovat käsikirjoitukset kulkeneet?

Mitä voi tehdä kokoteksteillä

- Tekstilouhinnan potentiaali on valtava. Esim. oikeudenmukaisuuden käsitteen muuttuminen brittiläisessä julkisessa keskustelussa: voitaisiin tutkia esimerkiksi louhimalla Early Modern English Books Online (EEBO) ja Eighteenth-Century Collections Online (ECCO) aineistoja, jotka yhdessä käsittävät käytännössä kaikki varhaisella uudella ajalla, 1470-1800, Britanniassa painetut kirjat.

- Käsitteen muutosta voidaan ryhtyä tutkimaan aivan eri tavoin kun käytössä on automatisoituja, laajojen tietoaineistojen käsittelyyn skaalautuvia mallintamismenetelmiä.

Vaivais-sanan käyttökontekstit 1820 – 1850



Useimmiten (91602 kertaa) Suomi24-keskustelupalstalla käytetty sana, jota automaattinen lauseenjäsennin <u>EI</u> tunnista:

Mitä voi tehdä kokoteksteillä ja metadatalla yhdessä

- Kontekstuaalinen lukija:
  - Ensimmäisen maailmansodan aikalaisteksteille
  - Antiikin teksteille
  - Suomalaiselle lakitekstille (2)
- Corpus of Early English Correspondence:
  - Käyttävätkö korkeasti koulutetut muita enemm onnellisuussanoja?

Tulevaisuus: Yhteismitallinen rahatiede

- Yhteistyössä kehitetty yhteismitallinen terminologia
- Paikkatunnisteet Pleiades-tietokannasta
- Avoimen lähdekoodin julkaisualusta joka julkaisee tiedon myös avoimena linkitettynä datana
- Useita julkaistuja palveluja:
  - Coin hoards of the Roman Republic Online
  - Coinage of the Roman Republic Online
  - Online Coins of the Roman Empire

Laajojen historiallisten aineistojen avoimen tutkijakäytön mahdollistamisen merkitys

- Innovatiivinen tutkimus on usein ruohonjuuritasolta nousevaa.
- Kehitystä ei tapahdu jos meillä on erikseen työkalujen tuottajat ja tutkijat.
- Parhaita käytäntöjä kannattaa lainata soveltuvasti muilta tieteenaloilta.
- Monitieteisissä projekteissa humanistisen ja laskennallisen puolen on kohdattava tavalla joka on syvempää kuin tehtävien jakaminen.
- Avoimuus sekä raaka- että tutkimusdatan kanssa on tie eteenpäin.

Tutkijoiden toivomus FinELibille

Olisi erittäin tärkeää, että kun ostetaan kalliita lisenssejä tekstitietokantoihin (ECCO, EEBO, NCCO jne.), niin sopimukseen pyritään lisäämään, että yliopiston tutkijoille ja tutkimusryhmille tarjotaan myös pääsy OCRttyyn raakatekstiin ja metadataan. Ilman sitä tiedonlouhintaan perustuva työ ei etene toivotulla tavalla.